(19) BUNDESREPUBLIK DEUTSCHLAND

DEUTSCHES PATENT- UND MARKENAMT

(12) Übersetzung der europäischen Patentschrift

(97) EP 0 689 189 B1

(10) DE 695 12 323 T 2

(51) Int. Cl.[7]: G 10 L 19/00
G 10 L 9/14
G 10 L 3/02
G 10 L 9/18

(21) Deutsches Aktenzeichen: 695 12 323.8
(96) Europäisches Aktenzeichen: 95 108 870.7
(96) Europäischer Anmeldetag: 8. 6. 1995
(97) Erstveröffentlichung durch das EPA: 27. 12. 1995
(97) Veröffentlichungstag der Patenterteilung beim EPA: 22. 9. 1999
(47) Veröffentlichungstag im Patentblatt: 6. 7. 2000

(30) Unionspriorität:
MI941283 20. 06. 1994 IT

(73) Patentinhaber:
Alcatel, Paris, FR

(74) Vertreter:
Dreiss, Fuhlendorf, Steimle & Becker, 70188 Stuttgart

(72) Erfinder:
Cucchi, Silvio, I-20090 Gaggiano (MI), IT; Fratti, Marco, I-20052 Monza (MI), IT

(84) Benannte Vertragstaaten:
DE, FR, GB, IT

(54) Methode zur Verbesserung der Eigenschaften von Sprachkodierern

DE 695 12 323 T 2

95 108 870.7-2215 /0 689 189 0132 010 fuh/neg 21.09.199
ALCATEL

1. BESCHREIBUNG DES STANDES DER TECHNIK

Sprachcodierung (Sprachverschlüsselung) wird in vielen Kommunikationsbereichen angewendet: von einer Übertragung über Satellit zum Mobilfunk, speichervermittelnden Systemen, automatische Antwortsender usw.

Insbesondere besteht ein starkes Bedürfnis nach wirksamen Techniken für die Sprachsignalcodierung dort, wo erkennbare Bandbegrenzungen vorhanden sind (betrachte die "begrenzte" Verfügbarkeit von Bandbreite in dem Äther); deshalb ist es wichtig, in der Lage zu sein, die zu übertragende Bitrate drastisch zu reduzieren und dabei weiterhin eine hohe Qualität des empfangenen Signals aufrecht zu erhalten.

Zu diesem Zweck werden verschiedene Sprachsignal-Codiertechniken verwendet; die üblichsten (die eine hohe Qualität des empfangenen Signals unter verschiedenen Bitraten sicherstellen) basieren auf den LP (Linear Prediction: lineare Vorhersage) und A-b-S (Analysis-by-Synthesis: Analyse durch Synthese)-Prinzipien (P. Kroon, E.F. Deprettere "A class of

analysis-by-synthesis predictive coders for high quality speech coding at rates between 4.8 and 16 Kbits/s", IEEE Journal on Selected Areas in Communications, Bd. 6, Nr. 2, Seiten 353-363, Februar 1988).

Die vorliegende Beschreibung offenbart einige Techniken zur Verbesserung der Eigenschaften von auf den vorstehend erwähnten Techniken basierenden Sprachcodierern. Nach einem Aspekt der Erfindung ist ein Anregungsparameter-Berechnungsverfahren gemäß Anspruch 1 angegeben. Nach einem weiteren Aspekt der Erfindung ist ein Toncodierer gemäß Anspruch 6 geschaffen.

Die Sprachcodierer, die auf der linearen Vorhersage (LP) basieren, sind parametrische Codierer; typischerweise werden Analyse-durch-Synthese-(A-b-S)-Techniken für eine korrekte Bestimmung der Parameter des Systems verwendet. Solche Codierer synthetisieren die Sprache durch die Verwendung einer geeigneten Eingangsanregung bei einem Synthese-LP-Filter.

Insbesondere sollte die Anregung die Charakteristiken der "physischen" Anregungssignalform aufweisen, die von der Stimmritze kommend dann als Funktion

der Charakteristiken des Systems, das das Sprachsegment simuliert (LP-Filter), spektral modifiziert wird.

Die modernsten A-b-S-Codierer verwenden eine Anregungsstruktur, die sich aus einem Adaptiven Codebuch und aus einem (eventuell strukturierten) Festen Codebuch zusammensetzen. Ohne Beeinträchtigung der Allgemeinheit kann angenommen werden, daß sich das Feste Codebuch aus unabhängigen Vektoren aus Zufallszahlen zusammensetzt, wie dies bei CELP-Codierern der Fall ist (M.R. Schroeder, B.S. Atal, "Code Excited Linear Prediction (CELP): high-quality speech at very low bit rates", Proc. ICASSP, '85, Seiten 937-940.

In Fig. 1 ist ein Blockdiagramm eines typischen CELP-Sprachsynthesizers dargestellt; Block LPC-IIR bezeichnet das Synthesizerfilter zur Rekonstruktion der Sprachsignalform; $e_a(n)$ ist der adaptive Codebuch-Vektor (und Ga ist der entsprechende Skalierungsfaktor) und $e_s(n)$ ist der feste Codelisten-Vektor (und Gs ist der entsprechende Skalierungsfaktor); e(n) ist der zusammengesetzte Anregungsvektor. Für eine detaillierte Beschreibung des Synthesizers kann auf W.B. Kleijn, D.J. Krasinski, R.H. Ketchum "Improved Speech Quality

and Efficient Vector Quantization in SELP", Proc. ICASSP '88, Seiten 155-158 Bezug genommen werden.

Im allgemeinen werden $e_a(n)$ und $e_s(n)$ aus einem geeigneten Satz von Vektoren gewählt und werden mit jeweiligen Ga und Gs gleichzeitig bestimmt. Die Bestimmung erfolgt in einem Zeitintervall von etwa 5 bis 10 ms (Analyserahmen) und basiert auf der Minimierung der Zielfunktion nach dem gut bekannten Kriterium des wahrnehmungsmäßig gewichteten quadratischen Mittelwertfehlers (siehe M.R. Schroeder, B.S. Atal, "Code Excited Linear Prediction (CELP): high-quality speech at very low bit-rates", Proc. ICASSP, '85, Seiten 937-940, gemäß dem folgenden Ausdruck:

$$E = \sum_{n=0}^{N-1} [r_s(n) - Gu_i(n)]^2 \qquad (1)$$

wobei N die Länge des Zeitintervalls für die Minimierung ist; $u_i(n)$ die Null-Zustand-Synthesefilterantwort an dem i-ten Eingang des Codebuches (entweder adaptiv oder fest) und G die entsprechende Verstärkung ist; schließlich ist $r_s(n)$ das Referenzsignal oder "Ziel"-Signal (d.h. das ur-

sprüngliche Sprachsegment, von dem der Beitrag des Rekonstruktionsfilterspeichers, abgeleitet von einer vorhergehenden Synthese, subtrahiert wurde).

Obgleich häufig verwendet, kann die bei (1) beschriebene Zielfunktion für die Wahl der Parameter nicht optimal sein. Insbesondere ist zu beachten, daß das System zufällig ist: dies bringt es mit sich, daß der von den Anregungsabtastungen in der Nähe von n = 0 herrührende Beitrag zu dem Synthesesignal im allgemeinen größer als der Beitrag ist, der von den Anregungsabtastungen in der Nähe von n = N - 1 herrührt. Diese Tatsache kann eine schlechte Näherung der idealen Anregung während Segmenten von Sprachsignalen bewirken. Unter diesen Umständen zeigt die ideale Anregung die Charakteristik von quasi-periodischen "Teilungsimpulsen". Diese synthetische Anregung soll in diesem Fall die Teilungsimpulse mit der richtigen zeitlichen Ausrichtung und der richtigen Amplitude beinhalten. In dem Fall, in dem sich die Impulse der idealen Anregung (üblicherweise als "Vorhersage-Rückstand" bezeichnet) an dem Ende des Minimierungsintervalls (d.h. für n in der Nähe von N - 1) befinden, wird ihre Rekonstruktion problematischer, da ihr Beitrag innerhalb des Minimierungsintervalls weniger "wiegt".

Dieses Phänomen wird während den Signaltransienten, d.h. in den Übergängen von sprachfreien Segmenten zu Sprachsegmenten und innerhalb der Sprachabschnitte in den Segmenten, in denen die ideale Anregung aufgrund der Vorhersagefiltervariationen ihre Form ändert (wobei weiterhin die "quasi-periodische" Charakteristik aufrechterhalten wird) noch deutlicher.

Im Folgenden werden zwei mögliche Vorgehensweisen zur Überwindung der vorstehend beschriebenen Probleme beschrieben; diese Vorgehensweisen können sowohl entweder einzeln als auch gemeinsam verwendet werden und ermöglichen, daß die Charakteristiken der bei verschiedenen Bitraten arbeitenden A-b-S-Codierer verbessert werden.

2. AUF FREIER ENTWICKLUNG BASIERENDE VORGEHENSWEISE

Eine erste Vorgehensweise besteht darin, als ein Referenzsignal der Zielfunktion (d.h. dem Signal $r_s(n)$ der Gleichung (1)) ein Signal $r_s^{el}(n)$ zu verwenden, das länger als N Abtastungen ist. Ein solches Signal wird aus der zeitlichen Verknüpfung der Signale $r_s(n)$ (für $n = 0 \ldots N - 1$) und aus der freien Entwicklung eines solchen Signals erhalten, und dieses freie Entwicklungs-el(n) wird erhalten, indem die letzten p Abtastungen von $r_s(n)$

in dem Synthesefilterspeicher LPC-IIR (wobei p die Ordnung des Filters ist) geladen werden und indem das Filter "entladen" wird, d.h. indem es seinen Ausgang entsprechend einem Null-Eingang berechnet.

Demzufolge wird erhalten:

$$r_s^{el}(n) = r_s(n), \quad n = 0..N-1 \qquad (2)$$

$$r_s^{el}(n) = el(n), \quad n = N..N-1+M \qquad (3)$$

wobei M die freie Entwicklungslänge ist.

Eine solche Vorgehensweise kann in folgender Weise gerechtfertigt werden: Die Sprache kann stets als von einer idealen Anregung erhalten betrachtet werden, was den Eingang eines Allpol-Synthesefilters (des in Fig. 1 mit LPC-IIR bezeichneten Filters) repräsentiert. Eine derartige ideale Anregung ist nichts anderes als die Vorhersageverzögerung, die durch eine Filterung der Sprache durch das "inverse Filter", d.h. das von LPC-IIR abgeleitete All-Null-(Dauer-Null)-Filter, erhalten wird.

Angenommen, man führt eine strichweise stationäre Analyse des Sprachsignals durch: Dann bildet die ideale Anregung innerhalb des Analyseintervalls den Zwangsausdruck für das Synthesefilter. Wenn jedoch

am Ende des Analyseintervalls der Eingang des Filters "ausgeschaltet" wird (d.h. die ideale Anregung auf Null gesetzt wird), wird das Synthesefilter gemäß einer Signalform entladen, die von seinen Polen und von den Abtastungen der idealen Anregung (insbesondere jenen, die dem Zeitpunkt n = N - 1 gerade vorhergehen) abhängt.

Es ist deshalb offensichtlich, daß in dem Fall, in dem die letzten Abtastungen der idealen Anregung wesentlich sind (beispielsweise wenn ein Tonteilungsimpuls vorhanden ist) und das Filter sich nahe einer Instabilität befindet (beispielsweise während Segmenten von Sprachsignalen), die freie Entwicklung des Filters aufgrund der idealen Anregung typischerweise sinusförmige Oszillationen zeigen wird, die ziemlich langsam abklingen werden und deshalb der Ausdruck el(n) der Gleichung (3) einen beträchtlichen Beitrag bilden wird.

Für eine hohe Qualität des rekonstruierten Signals ist es sehr wichtig, daß die synthetische Anregung spektrale und Zeitpunkt- (beispielsweise der Teilungsimpuls) Charakteristiken ähnlich jener der idealen Anregung hat. Es ist deshalb offensichtlich, daß durch Hinzunahme der Beiträge der sowohl auf die ideale Anregung als auch auf die syntheti-

sche Anregung zurückgehenden freien Entwicklungen in die Zielfunktion es möglich ist, eine korrektere Wahl der letzteren durchzuführen. Abhängig von den spektralen/zeitlichen Charakteristiken des Signals kann die Differenz zwischen der idealen freien Entwicklung und der synthetischen ein vorherrschendes Gewicht in der modifizierten Zielfunktion haben.

In Gleichungen können die vorstehend erwähnten Konzepte gemäß der umgeschriebenen Zielfunktion ausgedrückt werden:

$$El = \sum_{n=0}^{N-1+M} \left[r_s^{el}(n) - G u_i^{el}(n) \right]^2 \qquad (4)$$

in welcher

$$u_i^{el}(n) = u_i(n), \quad n = 0..N-1 \qquad (5)$$

$$u_i^{el}(n) = el_i(n), \quad n = N..N-1+M \qquad (6)$$

wobei $u_i(n)$ die (Null-Zustand)-Synthesefilterantwort an dem i-ten Eingang und $el_i(n)$ die entsprechende "synthetische" freie Entwicklung ist.

Die Anregungsparameter (d.h. der i-te Index und die entsprechende Verstärkung G) werden dann in solcher Weise gewählt, um die modifizierte Zielfunktion (4) zu minimieren.

Um beispielsweise die "ursprüngliche" freie Entwicklung el(n) zu erhalten, kann man in der folgenden Weise vorgehen:

- inverses Filtern (durch ein Sämtliche-Null-Filter) des Sprachsignals während des Intervalls 0...N - 1, wobei die ideale Anregung (Vorhersage-Rückstand), begrenzt auf das Zeitintervall 0 ...N - 1, erhalten wird.
- An dem Eingang des Synthesefilters LPC-IIR die dabei erhaltene ideale Anregung bereitstellen, und an dem Ausgang wieder das ursprüngliche Sprachsignal innerhalb des Zeitintervalls 0...N -1 erhalten.
- Ausgehend von dem auf diese Weise erhaltenen Endstatus des Synthesefilters Bereitstellen eines Nulleingangs an dem Eingang des Synthesefilters, und das Filter für eine Anzahl M von Abtastungen gleich der Länge der zu erhaltenden freien Entwicklung "entladen" lassen.

Aus der vorstehend beschriebenen Prozedur wird sofort ersichtlich, daß keine Notwendigkeit besteht,

den Vorhersage-Rückstand zu berechnen. Um die gewünschte freie Entwicklung zu erhalten, ist es ausreichend, die letzten p Abtastungen (p stellt die Ordnung des Filters dar) des ursprünglichen Sprachsignals (d.h. der Abtastungen N - 1, N - 2, ..., N - p) in den Zustand des Synthesefilters zu zwingen und das Filter mit Nulleingang entladen zu lassen. Offensichtlich kann man für die Berechnung der synthetischen freien Entwicklung in ähnlicher Weise vorgehen.

Schließlich ist zu beachten, daß diese Vorgehensweise keine Zunahme der Codierverzögerung mit sich bringt, da in der Zielfunktion die Sprachabtastungen jenseits des Zeitintervalls 0...N - 1 nicht verwendet werden.

3. DIE GEWICHTUNGS-BASIERENDE VORGEHENSWEISE

In dem vorhergehenden Abschnitt wurde dargelegt, daß es zur Erzielung einer hohen Qualität des rekonstruierten Signals sehr wichtig ist, daß die synthetische Anregung spektrale und Zeitpunkt-(beispielsweise Teilungsimpuls)-Charakteristiken aufweist, die ähnlich zu jenen sind, die bei der idealen Anregung vorliegen. Daraus folgt, daß es wichtig sein kann, nicht nur eine gute Ähnlichkeit zwischen der ursprünglichen Sprache und der syn-

thetischen Sprache zu erhalten, sondern auch eine gute Übereinstimmung zwischen der idealen Anregung und der synthetischen Anregung zu erhalten.

Durch Verwendung einer Vorgehensweise der minimalen Quadrate in der klassischen Zielfunktion ermöglichen es die Parameter der rekonstruierten Anregung tatsächlich, eine synthetische Sprache zu erzielen, die "im Durchschnitt" ähnlich zu der ursprünglichen Sprache ist.

Unter dem Gesichtspunkt der Wahrnehmung ist es tatsächlich manchmal wichtiger, daß die synthetische Sprache nur lokal der ursprünglichen Sprache ähnlich ist (beispielsweise ist es sehr wichtig, die Verbindung von einem sprachfreien Segment zu einem Sprachsegment innerhalb der richtigen zeitlichen Ausrichtung und mit der korrekten Dynamik zu rekonstruieren. Es ist nicht ungewöhnlich, Verbindungstransienten zu finden, deren Zeitdauer sehr viel kürzer als die Zeitdauer des Syntheserahmens ist). Dann ist es für eine ziemlich lokale Rekonstruktion wichtig, einen gewissen Grad an Ähnlichkeit auch mit der idealen Anregung aufrecht zu erhalten.

Die Zielfunktion kann sich dann aus zwei Beiträgen, als Funktion der ursprünglichen Sprache bzw. der idealen Anregung, zusammensetzen und nimmt den folgenden Ausdruck an:

$$E2 = \alpha E + (1 - \alpha) E3 \qquad (7)$$

wobei:

$$E = \sum_{n=0}^{N-1} [r_s(n) - Gu_i(n)]^2 \qquad (8)$$

$$E3 = \sum_{n=0}^{N-1} [e_s(n) - Ge_i(n)]^2 \qquad (9)$$

In Gleichung (9) ist $e_s(n)$ der von dem Referenzsignal $r_s(n)$ erhaltene Vorhersage-Rückstand und $e_i(n)$ ist die Codebuch-Anregung, die das synthetische Signal $u_i(n)$ erzeugt. Es ist zu beachten, daß der Vorhersage-Rückstand $e_s(n)$ ausgehend von $r_s(n)$ durch eine inverse Filterung (mit einem Sämtlich-Null-Filter) mit einem ursprünglichen Null-Zustand berechnet werden muß. Wie bekannt ist, wurde die Referenz tatsächlich aus dem Sprachsignal durch Subtraktion ihrer Rekonstruktionsfilter-Spei-

cherermittlung, abgeleitet von der vorhergehenden Synthese, erhalten. Das Referenzsignal ist dann "frei" von jeglichem auf den Filterspeicher zurückgehenden Beitrag und kann als von einer geeigneten idealen Anregung $e_s(n)$ erhalten betrachtet werden, die mit einem anfänglichen Null-Zustand in das Synthesefilter einläuft.

In Gleichung (7) ist α ein Parameter, dessen Wert zwischen 0 und 1 liegt und die Bedeutung steuert, die der Minimierung im Hinblick auf das Referenzsignal zukommt. Bei $\alpha = 1$ wird die ursprüngliche Zielfunktion wieder erhalten.

Die Anregungsparameter (d.h. der i-te Index und die entsprechende Verstärkung G) werden dann derart gewählt, um die in den Gleichungen (7), (8), (9) beschriebene Zielfunktion zu minimieren. Der Parameter α kann entweder fest oder sogar adaptiv (d.h. zeitlich variierend) sein, beispielsweise als Funktion bestimmter Charakteristiken des Signals, das a priori geschätzt werden kann (beispielsweise: Schätzung von sprachbehaftet/sprachfrei, Schätzung der Transienten, Schätzung der Teilungsperiode oder des Synthesefilters, usw.).

Schließlich ist zu beachten, daß der in dem vorhergehenden Abschnitt beschriebene, auf die freie Entwicklung zurückgehende Beitrag in der durch die Gleichungen (7), (8), (9) beschriebenen Zielfunktion mit einbezogen werden kann. In diesem Fall wird der Ausdruck (8) der Zielfunktion gemäß der Beschreibung in dem vorhergehenden Abschnitt modifiziert.

95 108 870.7-2215/0 689 189 0132 010 fuh/neg 21.09.1999
Anmelder: ALCATEL

P a t e n t a n s p r ü c h e

1. Verfahren zum Berechnen der Anregungsparameter in Sprachcodierern basierend auf linearen Vorhersage- und Analyse-durch-Synthese-Techniken, die eine zu minimierende Zielfunktion verwenden, **dadurch gekennzeichnet**, daß die Zielfunktion gemeinsam oder alternativ a) die freie Entwicklung des Zielsignals und des synthetischen Signals und b) eine Gewichtung im Hinblick auf den Fehler zwischen dem Vorhersage-Rückstand und der synthetischen Anregung umfaßt.

2. Verfahren nach Anspruch 1 in den Alternativen a) oder a) und b), dadurch gekennzeichnet, daß die Zielfunktion:

$$Ex = \alpha E1 + (1 - \alpha) E3 \qquad (10)$$

verwendet wird, wobei die Funktion E1 neben dem Fehler zwischen den Zielsignalen und den synthetischen Signalen auch den Fehler zwi-

schen den relativen freien Entwicklungen berücksichtigt, und die Funktion E3 den Fehler zwischen dem Vorhersage-Rückstand und der synthetischen Anregung berücksichtigt, und $0 < \alpha \leq 0$ ist.

3. Verfahren nach Anspruch 2, dadurch gekennzeichnet, daß die Funktion E1 gegeben ist durch:

$$E1 = \sum_{n=0}^{N-1+M} \left[r_s^{el}(n) - Gu_i^{el}(n) \right]^2 \qquad (11)$$

wobei N die Länge des Zeitintervalls für die Minimierung ist, M die freie Entwicklungslänge ist, $r_s^{el}(n)$ das durch eine freie Entwicklung erhaltene erweiterte Referenzsignal ist, $u_i^{el}(n)$ die erweiterte Null-Zustands-Synthesefilterantwort an dem i-ten Eingang des Codebuches ist, und G die entsprechende Verstärkung ist.

4. Verfahren nach Anspruch 2, dadurch gekennzeichnet, daß die Funktion E3 gegeben ist durch:

$$E3 = \sum_{n=0}^{N-1} [e_s(n) - Ge_i(n)]^2 \qquad (12)$$

wobei $e_s(n)$ der von dem Referenzsignal erhaltene Vorhersage-Rückstand ist und $e_i(n)$ das Codebuch-Anregungssignal ist.

5. Verfahren nach Anspruch 2, dadurch gekennzeichnet, daß der Gewichtsfaktor zeitlich variierbar ist.

6. Toncodierer, der umfaßt:
Mittel zum Ausführen einer linearen Vorhersage,
Mittel zum Ausführen einer Analyse-durch-Synthese, und
Mittel zum Berechnen der Anregungsparameter unter Verwendung einer zu minimierenden Zielfunktion,

dadurch gekennzeichnet, daß die Zielfunktion gemeinsam oder alternativ

a) die freie Entwicklung des Zielsignals und des synthetischen Signals, und

b) eine Gewichtung im Hinblick auf den Fehler zwischen dem Vorhersage-Rückstand und der synthetischen Anregung umfaßt.

22.09.99

1/1

95 108 870.7-2215/0 689 189 0132 010 fuh/neg 21.09.1999
Anmelder: ALCATEL

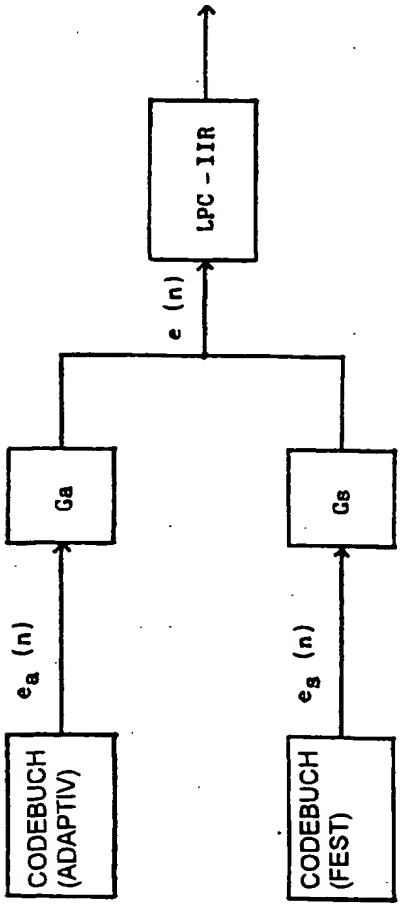

FIG. 1

This Page Blank (uspto)